Selection on Codon Bias

Ruth Hershberg and Dmitri A. Petrov

Department of Biological Sciences, Stanford University, Stanford, California 94305; email: dpetrov@stanford.edu

Annu. Rev. Genet. 2008. 42:287–99

The *Annual Review of Genetics* is online at genet.annualreviews.org

This article's doi: 10.1146/annurev.genet.42.110807.091442

0066-4197/08/1201-0287$20.00

Key Words

codon bias, selection, evolution

Abstract

In a wide variety of organisms, synonymous codons are used with different frequencies, a phenomenon known as codon bias. Population genetic studies have shown that synonymous sites are under weak selection and that codon bias is maintained by a balance between selection, mutation, and genetic drift. It appears that the major cause for selection on codon bias is that certain preferred codons are translated more accurately and/or efficiently. However, additional and sometimes maybe even contradictory selective forces appear to affect codon usage as well. In this review, we discuss the current understanding of the ways in which natural selection participates in the creation and maintenance of codon bias. We also raise several open questions: (*i*) Is natural selection weak independently of the level of codon bias? It is possible that selection for preferred codons is weak only when codon bias approaches equilibrium and may be quite strong on genes with codon bias levels that are much lower and/or above equilibrium. (*ii*) What determines the identity of the major codons? (*iii*) How do shifts in codon bias occur? (*iv*) What is the exact nature of selection on codon bias? We discuss these questions in depth and offer some ideas on how they can be addressed using a combination of computational and experimental analyses.

THE CODON BIAS PHENOMENON

The genetic code determines which of the 61 triplets or codons correspond to which of the 20 amino acids. Because there are more codons than amino acids, the genetic code is necessarily redundant. While few amino acids are encoded by a single codon, most amino acids are encoded by two to six different codons. Different codons that encode the same amino acid are known as synonymous codons. Changes in the DNA sequence of a protein between two synonymous codons are often assumed to have no effect and are thus called synonymous changes or even silent changes. However, even though synonymous codons encode the same amino acids, it has been shown for a wide variety of organisms that different synonymous codons are used with different frequencies. This phenomenon has been termed codon bias.

The genetic code is generally conserved among organisms, but the direction of codon bias shifts between different organisms. Thus the identity of the more and less frequent codons for each amino acid differs between organisms. At the same time, the choice of which codons are frequent and which are rare is generally consistent across genes within each genome (11, 21, 24). The hypothesis that different organisms have distinct codon biases is known as the genome hypothesis of codon bias (21).

Strength of codon bias also varies between organisms. In some organisms codon bias is very strong, whereas in others the different synonymous codons are used with similar frequencies (5, 6, 14, 24, 29, 30, 38, 40, 41). Likewise, the strength of codon bias varies across genes within each genome, with some genes using a highly biased set of codons and others using the different synonymous codons with similar frequencies (20, 24, 39).

POSSIBLE EXPLANATIONS FOR CODON BIAS: MUTATIONAL BIAS AND/OR SELECTION

Explanations for the existence of codon bias fall into two general classes (7, 14, 40, 41). According to the selectionist explanation, codon bias contributes to the efficiency and/or the accuracy of protein expression and is thus generated and maintained by selection. The mutational or neutral explanation, by contrast, posits that codon bias exists because of nonrandomness in the mutational patterns. Some codons are more mutable and thus would have lower equilibrium frequencies. Mutational biases are known to differ between organisms, possibly leading to differences in the patterns of codon bias across organisms.

Corroboration of the mutational explanation of codon bias can be seen in several studies that have shown that the most significant parameter explaining codon bias differences between different organisms is the level of GC content (11, 25, 28). GC content is likely to be determined mostly by genome-wide processes rather than by selective forces acting specifically on coding regions. In fact, one study demonstrated that the differences in codon bias among prokaryotes may be predicted by using statistics gleaned solely from intergenic sequences (11).

The above finding is consistent with the genome-wide patterns of codon usage being determined by mutational biases. However, there are some clear indications that natural selection must also be involved. Mutational pressures alone cannot explain why the more frequent codons (also called preferred codons) are those that are recognized by more abundant tRNA molecules (24, 26, 27, 46). This correlation was detected through direct measurements of tRNA levels in the bacteria *Escherichia coli* and *Mycoplasma capricolum* and in the yeast *Saccharomyces cerevisiae* (24, 46). It has also been detected based on the tRNA gene copy numbers (which have been shown in *E. coli*, *Bacillus subtilis*, and yeast to correlate with cellular tRNA abundances) in many additional bacteria (27), as well as in several eukaryotic species (*Schizosaccharomyces pombe*, *Drosophila melanogaster*, and *Caenorhabditis elegans*) (26).

The mutational model also does not easily account for within-genome variation in codon bias. Codon bias correlates most strongly with the level of gene expression (14, 20, 24).

Correlations between levels of gene expression and codon bias have been shown using large-scale gene expression data, in organisms as diverse as *E. coli*, *S. cerevisiae*, *C. elegans*, *Arabidopsis thaliana*, and *D. melanogaster* (10, 15, 18, 19). It has also been demonstrated in a large number of bacteria and in yeast that genes that interact functionally and thus likely need to be expressed at similar levels tend to have correlated levels of codon bias (17, 31).

In principle, the relationship between codon bias and gene expression may be due to differences in mutational biases in genes transcribed at different levels (16). However, studies in *D. melanogaster* and *C. elegans* suggest that this is unlikely (14, 15). In both of these organisms most of the optimal codons contain a cytosine or a guanine in the third position. As a result, the GC content of synonymous sites in these organisms correlates positively with levels of gene expression. However, the same is not true for the GC content of introns that are also transcribed and should be affected by the same transcription-coupled mutational processes as synonymous sites (14, 15).

Both the match between more commonly used codons and the abundant tRNAs, as well as the high codon bias of the highly expressed genes, fit well with the selectionist explanation for codon bias. Genes using the codons that are recognized by more abundant tRNA molecules may be translated more efficiently and with fewer mistakes than genes that use less frequent codons. Thus, selection may favor the use of the more frequent codons. Such selection is expected to be stronger for genes that are expressed at higher levels, fitting well with the observed correlation between levels of gene expression and levels of codon bias.

The exact cause of selection for translationally optimal codons is unclear. Selection may be due to the need to maximize the speed of elongation and to increase the cellular concentration of free ribosomes and/or to minimize the incorporation of the wrong amino acids into the nascent polypeptide chain. Additional selective pressures that may have little to do with translation, may also be in play.

An early study by Precup & Parker (36) looked at the frequency of misincorporation of lysine into the AAU and AAC asparagine codons by experimentally constructing a series of derivatives of the gene encoding the coat protein of the bacteriophage MS2 (36). This study showed that the choice of a codon strongly affected (four- to ninefold) the frequency of misincorporation (36). Studies conducted by Akashi (1) and by Stoletzki & Eyre-Walker (44) further demonstrated that selection on translational accuracy seems to play a role in codon bias in *D. melanogaster* and *E. coli*. Both studies used a test originally suggested by Akashi (1); if selection on codon bias acts to increase translational accuracy, it should act more strongly on codons that encode the functionally most important amino acids. The authors found that this was indeed the case and that the sites that encode more conserved amino acids are also more biased in terms of codon usage (1, 44).

Codon usage can also affect the speed of translation elongation. Curran & Yarus (12) observed that the rate for aminoacyl-tRNA selection by different codons spans a 25-fold range and that preferred codons select their aminoacyl-tRNAs more quickly than more rarely used codons (12). Sorensen et al. (43) measured elongation rates directly in vivo. They showed that the insertion of short strings of either rare codons or frequent codons significantly affects the rate of elongation. The rate of amino acid incorporation at the frequent codons was almost six times faster than the rate at the rare codons (43). However, whether the increased speed of translation elongation at preferred codons is advantageous is not clear (7). One possibility is that proteins that are encoded by more frequent codons can be translated more quickly. However, it has not yet been determined whether increased speed of elongation will lead to a noticeable increase in the speed of translation. This will be true only if elongation rather than initiation of translation is the limiting step for polypeptide biosynthesis. Some evidence indicates that this is not the case (7). However, even if increasing the elongation rate does not directly increase the speed of

translation for a particular gene, it may still increase the pool of free ribosomes and thus indirectly increase the rate of initiation for all messenger RNAs. This should increase the rate of translation for all proteins and is likely to be advantageous overall.

Carlini & Stephan (8, 9) manipulated the sequence of the alcohol dehydrogenase (*Adh*) gene in *Drosophila* by replacing 1 to 10 preferred codons with unpreferred ones while maintaining the amino acid sequence of the protein. They showed that these relatively small changes in codon bias affect both the expression of the ADH gene as well as the ability of the flies carrying the altered genes to tolerate ethanol (8, 9). These studies directly demonstrate the strong effect codon bias may have on gene function and possibly on fitness.

Given the evidence that both mutational pressures and selection are involved in the phenomenon of codon bias, the current accepted model is the major codon preference model, also known as the mutation-selection-drift balance model of codon bias (2–4, 7, 14). This model proposes that selection favors the major (or preferred) codons over minor codons. However, mutational pressure and genetic drift allow the minor codons to persist. Factors such as levels of gene expression and functional constraint may determine the intensity of selection on silent sites for a certain gene. For example, selection on silent sites in ribosomal genes that are more highly expressed and more functionally constrained may be stronger than selection on the silent sites of less constrained and/or less highly expressed genes. Under this mutation-selection-drift balance model, codon bias is the result of positive selection for mutations that increase the frequency of major codons (preferred mutations) and purifying selection against mutations that decrease the frequency of major codons (unpreferred mutations). This model postulates that the selection on codon bias is generally weak. In the next section, we discuss the population genetics evidence in support of the mutation-selection-drift balance model of codon bias.

POPULATION GENETICS STUDIES INTO THE MUTATION-SELECTION-DRIFT BALANCE MODEL

Population genetics approaches have been used to establish whether the major codon preference model can explain codon bias. If codon bias is indeed subject to selection, differences in the evolutionary dynamics of different classes of synonymous mutations would be expected. Specifically, mutations from a minor codon to a major codon should be beneficial, whereas mutations in the opposite direction should be deleterious. This difference should influence the fate of the preferred and unpreferred mutations. The probability that a mutation will either decrease or increase in frequency depends on the product of the effective population size (N_e) and the selection coefficient (s). When compared to what we would expect for neutrally evolving sequences, positive selection ($s > 0$) should elevate the number of fixed differences between species and the frequencies with which these sites segregate within a population or species. Purifying selection ($s < 0$) should have opposite effects (22).

McDonald & Kreitman (33) compared the ratios of the number of segregating sites within a population to the number of fixed differences between species for different classes of mutations. This ratio is often referred to as r_{pd}. r_{pd} decreases in a monotonic fashion as $N_e s$ moves from negative to positive values. Originally, McDonald & Kreitman compared the r_{pd} ratios for nonsynonymous and synonymous changes at the *Adh* locus of *Drosophila* (33). This test can be used to investigate whether selection is acting on codon bias by comparing different classes of synonymous changes. If selection is indeed acting on synonymous mutations as predicted by the mutation-selection-drift balance model, we would expect r_{pd} to be higher for unpreferred changes than for preferred changes. Moreover, deleterious mutations should segregate at lower frequencies than neutral mutations, whereas advantageous mutations should segregate at higher frequencies. It is thus

possible to compare the frequencies with which preferred and unpreferred synonymous changes segregate within a population and to examine whether preferred mutations do indeed segregate at higher frequencies.

The seminal studies using these population genetic approaches to corroborate the major codon preference model were conducted in *Drosophila* by Hiroshi Akashi (2–4). Akashi first identified the *Drosophila* major codons as those codons that increase in frequency as a function of the calculated codon bias for all other amino acids. He then examined the evolution of codon bias since the split of the closely related *Drosophila* species, *D. melanogaster* and *D. simulans* (2). At the time Akashi conducted his studies only a limited amount of sequence data was available. He analyzed five genes for which at least two alleles were sequenced in both *D. melanogaster* and *D. simulans* and in at least one other species within the *D. melanogaster* subgroup. Consistent with the major codon preference model, he found that r_{pd} is significantly higher for unpreferred than for preferred mutations along the *D. simulans* lineage (2). For *D. melanogaster* the r_{pd} values for the two classes of synonymous mutations did not differ significantly. Akashi thus suggested that selection at silent sites is less effective in *D. melanogaster* than in *D. simulans* (2). He also pointed to the fact that synonymous fixation rates for unpreferred mutations are higher in *D. melanogaster*, which may confirm a genome-wide relaxation of selection at silent sites in this lineage. Akashi suggested that this may be the result of differences in the effective population sizes of *D. melanogaster* and *D. simulans* (2). Such a difference in N_e could explain the reduction in selection on codon bias observed in *D. melanogaster*. Note that we would expect such a sensitivity to effective population size only in the case of weak selection ($|N_e s| \approx 1$).

Akashi further used the Poisson random field method of Sawyer & Hartl (37) to estimate N_es for synonymous mutations in *D. simulans*, based on the r_{pd} values for two of the genes analyzed. He determined that the value of N_es for these genes ranged between –3.6 and –1.3, i.e., in the range of weak selection (2).

In a second study, Akashi, together with Schaeffer (4), tested whether preferred mutations segregate at higher frequencies than unpreferred mutations. They analyzed nine genes for which multiple alleles had been sequenced in both *D. melanogaster* and *D. simulans* and which had been sequenced in at least one outgroup species within the *D. melanogaster* subgroup. The authors also examined 99 alleles of the two genes contained in the Adh region of *D. pseudoobscura* (*Adh* and *Adhr*).

Major codons were defined as in the previous Akashi study (2), and the direction of synonymous mutations was again decided using parsimony and outgroup sequences. The frequency spectra of preferred and unpreferred synonymous mutations were examined in the three *Drosophila* species. However, there were not enough preferred mutations segregating among the six alleles of the nine genes examined in *D. melanogaster* and so the results could be analyzed only for the other two *Drosophila* species. Significant differences could be seen in both *D. pseudoobscura* and *D. simulans* in the frequencies of preferred and unpreferred synonymous mutations. Fitting with a model of major codon preference, preferred mutations were found to segregate at significantly higher frequencies than unpreferred mutations (4).

The Akashi studies were key in showing that selection does indeed affect the silent sites of proteins. However, by necessity they used only a small amount of data. As the sequence data became more readily available in a large number of species, McVean & Vieira (34) devised a method that relied on the combination of population genetic models and likelihood methods of phylogenetic sequence analysis to estimate parameters of both mutation and selection. They compared 50 orthologous gene pairs from *D. melanogaster* and *D. virilis* and 27 from *D. melanogaster* and *D. simulans*. Their method and the increased size of their dataset allowed them to show that the strength of selection on codon bias varies considerably between different amino acids and different

genes. They also showed considerable variation in the strength of selection between different *Drosophila* species (34). Most remarkably, *D. melanogaster* showed no evidence of current selection on codon bias (34).

More recently, Nielsen et al. (35) presented a likelihood method for estimating codon bias parameters similar to that of McVean & Vieira but that can be applied to more than two species. In addition, the Nielsen method uses a more complex mutation model than McVean and Vieira's, one that allows for differences in mutation pressures for different lineages. Their method is an extension of the popular maximum likelihood methods used to estimate dN/dS ratios along the branches of a phylogenetic tree that are implemented in the commonly used software package PAML (47, 48). They extend these methods by adding a parameter that represents the selection coefficient for optimal codon usage for each branch of the tree. Thus, they can use their method to simultaneously estimate mutation rates, dN/dS, and the strength of selection acting on codon bias, along each branch of a phylogenetic tree. An initial application of this method using 18 genes indicated that differences exist both in mutation rate and in the strength of selection on codon bias between *D. melanogaster* and *D. simulans*. In *D. simulans*, the results supported the major codon preference model. However, in *D. melanogaster* only 1 of the 18 genes, *Notch*, showed evidence of selection on synonymous sites (35). Interestingly and consistent with the results of a previous study by DuMont et al. (13), *Notch* appears to have been evolving in the *D. melanogaster* lineage under selection in favor of unpreferred codons. This finding clearly does not fit the major codon preference model but rather points to the possibility that other selective forces, in addition to major codon preference, may be acting on synonymous sites.

To further explore the patterns of synonymous site evolution in *Drosophila*, Singh et al. (42) applied the Nielsen algorithm (35) to 8452 genes with clear orthologs in *D. melanogaster*, *D. sechellia*, and *D. yakuba*. The authors found that in *D. melanogaster* there was little evidence for recent selection on synonymous sites. The *Notch* gene was again found to be an outlier. In *D. sechellia*, by contrast, selection was acting predominantly in favor of preferred codons (42). However, even in this species, for a small number of genes selection seems to favor the unpreferred codons, which indicates that sometimes selection may be acting on codon bias for reasons other than to enhance the efficiency or accuracy of translation. In agreement with previous studies, the authors estimated that the strength of selection on synonymous sites in *D. sechellia* is quite weak. The median N_es for genes under selection in favor of preferred codons was 2.04 (42).

In a very recent study, Yang & Nielsen (49) adjusted Nielsen's method to relax some of its assumptions. For instance, the original method required a priori assignment of synonymous codons into preferred and unpreferred. The authors applied their algorithm to test for selection on codon bias in five mammalian species: human, chimpanzee, macaque, mouse, and rat. They found that in most genes, they could reject the null hypothesis that codon bias is due only to mutation bias and is not influenced by selection. This may suggest that even in mammals selection is affecting the evolution of codon bias. This is important because the phenomenon of codon bias is much weaker in vertebrates and the effect of selection on codon bias in vertebrates is widely disputed (14). Yang & Nielsen estimated that selection on codon bias in mammals, albeit significant, is weak and that most of the synonymous mutations are nearly neutral (49).

OPEN QUESTIONS

Is Selection on Codon Bias Constant and Weak?

Under the major codon preference model, codon bias is maintained by mutation-selection-drift balance. Selection increases the usage of preferred codons while mutation counters this increase. Consider a codon family with only two codon states, preferred and

unpreferred. Let the mutation rates from preferred to unpreferred and from unpreferred to preferred be μ_p and μ_u, respectively. At equilibrium the proportion of codons that are fixed for the preferred state should be (42a):

$$P(preferred) = \frac{1}{\left(1 + \frac{\mu_p}{\mu_u} * e^{-4Nes}\right)},$$

where N_e is the effective population size and s is the strength of selection for preferred codons.

For most genes, levels of codon bias are intermediate, meaning that the proportion of codons fixed for the preferred state is intermediate for most genes. It is easy to see, looking at the above equation, that s has to be within a very limited range in order for the proportion of codons that are fixed for the preferred state to be intermediate ($s \sim 1/N_e$). Thus selection on synonymous sites must clearly be weak.

The studies described in the previous section implied but never clearly stated that the selection on synonymous sites is constant. Such a model of constant selection can be described by a linear function of the form Fitness = s*[level of codon bias] + constant (**Figure 1**). On this model, for most genes s is a constant on the order of $1/N_e$. Because most genes show intermediate levels of codon bias both in *E. coli* and *Drosophila* and because these organisms likely have very different effective population sizes, s must somehow be very precisely negatively correlated with the effective population size. It is unclear why this would be the case.

Finally, the experiments showing that changes to as few as 10 synonymous codons in the *Adh* gene of *Drosophila* greatly affect both the expression of the ADH protein and the ability of *Drosophila* to tolerate ethanol seem to indicate that changes in codon bias can have noticeable effects on the biology of the organism (8, 9). Although not necessarily inconsistent with weak selection, these findings hint at the possibility that selection acting on codon bias might be strong, at least in some cases.

An alternative possibility is that codon bias evolves to a level at which selection on synonymous sites is weak. In this model, selection is weak when codon bias reaches the equilibrium value but might be quite strong when codon bias is far from the equilibrium value (**Figure 2**). Different genes vary in their optimal levels of gene expression. For each level of gene expression some levels of codon bias are unacceptably low and generate strong negative fitness effect. This generates a strong selective pressure to elevate codon bias. As the

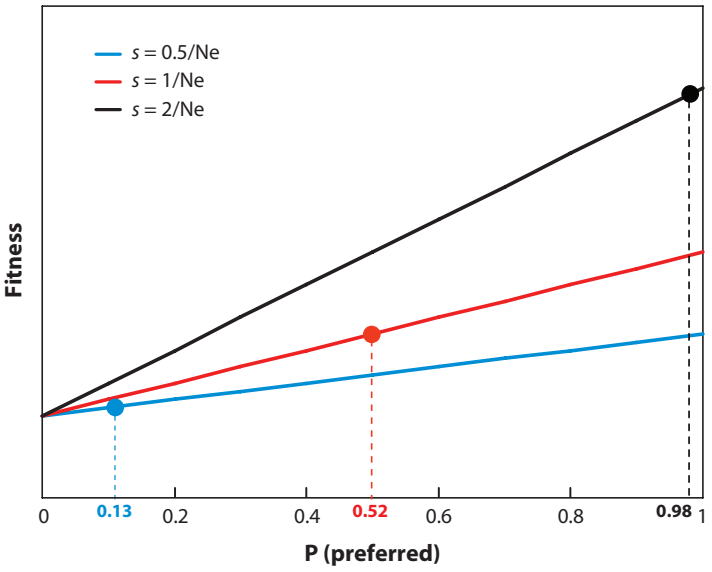

Figure 1

The relationship between codon bias and fitness under a model of constant selection on synonymous sites. Under this model, fitness is a linear function of the level of codon bias, so that fitness = s* (level of codon bias) + constant, where s is the selection coefficient and (level of codon bias) is the proportion of codons that are fixed for the preferred state. The relationship between fitness and p(preferred) is drawn for three different selection coefficients ($s = 0.5/N_e$, $s = 1/N_e$, and $s = 2/N_e$). For this figure $N_e = 10^6$, constant = 0.2. The equilibrium proportion of codons that are fixed for the preferred state (*small circles*) was calculated for each value of s using the formula:

$$\hat{P}(preferred) = \frac{1}{\left(1 + \frac{\mu_p}{\mu_u} * e^{-4Nes}\right)},$$

where μ_p and μ_u are the mutation rates from preferred to unpreferred and from unpreferred to preferred, respectively. For this calculation the ratio between the two was arbitrarily set to 50. For the equilibrium proportion of codons that are fixed for the preferred state to be intermediate (as is true for most genes), the selection coefficient, s, must be in the range of $1/N_e$, as larger values of s lead to genes that are almost entirely encoded by preferred codons (exemplified here by the case of $s = 2/N_e$), whereas lower values of s lead to genes with levels of codon bias determined solely by mutation pressures (exemplified here by the case of $s = 0.5/N_e$).

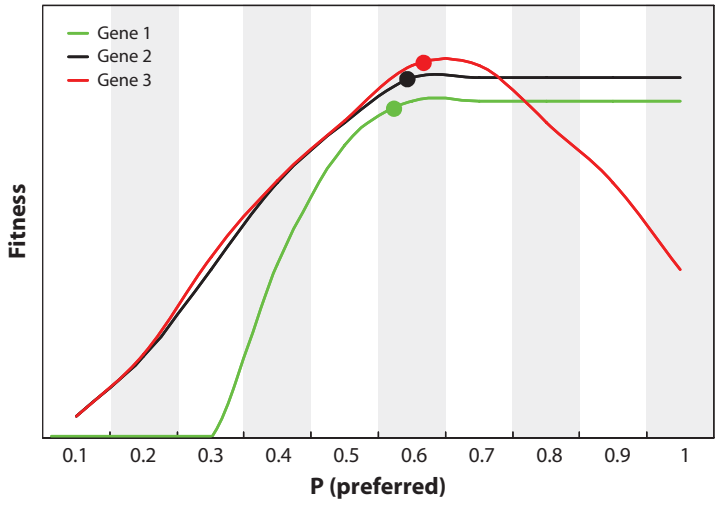

Figure 2

Schematic representation of possible nonlinear, nonconstant relationships between fitness and codon bias. The strength of selection changes as a function of codon bias. At equilibrium, selection is weak and a regimen of selection-mutation-drift is reached. The equilibrium proportion of codons that are fixed for the preferred state are marked by small circles. Three possible relationships between fitness and codon bias are depicted: For all three genes at low levels of codon bias, selection for preferred mutations is strong and the strength of selection decreases as codon bias nears equilibrium. For gene 1 very low levels of codon bias are lethal. For both gene 1 and gene 2 fitness reaches a plateau after which additional preferred mutations neither increase nor decrease fitness. For gene 3 there is an optimal level of codon bias that is close to the equilibrium level of codon bias for that gene, and once the gene reaches this level, additional preferred mutations are selected against.

codon bias increases, the strength of selection toward additional preferred changes decreases until the equilibrium level of codon bias is reached. At the equilibrium level there is a regimen of selection-mutation-drift balance.

The shape of the function describing the relationship between the strength of selection in favor of preferred synonymous changes(s) and codon bias may differ between genes. This in turn will change the relationship between codon bias and fitness. In **Figure 2** we show three possible scenarios for the relationship between fitness and codon bias for three theoretical genes with similar expression levels. For both gene 1 and gene 2, fitness increases until the gene reaches a certain level of codon bias. However, there is no fitness cost or benefit to increasing codon bias beyond that level. For gene 3, there is an optimal level of codon bias and increasing codon bias beyond this level reduces fitness. In gene 1, very low levels of codon bias result in lethality. This may be the case for some highly expressed genes. For such highly expressed genes, low codon bias might result in both massive shortage of ribosomes as well as severe accumulation of mistranslated and perhaps misfolded proteins.

It is possible to use a combination of sequence and experimental analyses to examine directly whether selection on preferred and unpreferred changes is constant or not. Under a model of constant and weak selection, a change in codon bias does not affect the strength of selection applied to subsequent mutations. However, if the strength of selection changes as a function of codon bias, the fate of synonymous mutations should be affected by the synonymous mutations that preceded them. It may be possible to examine this by following the occurrence of preferred and unpreferred synonymous changes along a phylogenetic tree.

It would also be useful to experimentally examine the effects of incorporating preferred and unpreferred mutations that have occurred along the different branches of the tree out of the order at which they naturally occurred. Such an analysis might allow for direct probing of the relationship between the strength of selection on preferred and unpreferred synonymous mutations and current levels of codon bias.

If the selection acting on synonymous sites is strong when codon bias is far from its equilibrium value, horizontally transferred genes that have a codon usage divergent from that of their new host genome may be able to persist only if they are not expressed at very high levels. Sharp increases in transcription levels without corresponding increases in codon bias might be strongly deleterious in general. It is tempting to speculate that increases in the expression levels of genes are achieved by the stepwise process of increasing the efficiency of transcription and translation: A small increase in the transcription of the gene will increase selection in favor of preferred synonymous changes. This will increase codon bias, which will allow for a further increase in transcription efficiency, thereby

further increasing the strength of selection in favor of preferred synonymous mutations. The final result of this iterative process should be a highly transcribed and highly codon-biased gene.

What Determines the Choice of Major Codons?

As noted above, differences in codon bias between organisms can be predicted based on the composition of intragenic sequences (11). This is seen as proof that interspecies differences in the direction of codon bias are driven largely by differences in genome-wide patterns of substitution. This, however, explains only the codon usage in the genes showing low levels of codon bias within the genome. Still unclear is what determines which of the synonymous codons will be used as major and minor codons. In some cases, the choice of major codons appears to be strongly nonrandom and thus hard to explain. For example, in *Drosophila* and in *C. elegans* most major codons end in either a cytosine or a guanine even though both genomes are AT rich (14). No theories have been proposed, to our knowledge, to explain this fact.

How Do Shifts in Codon Bias Occur?

It is unclear how shifts in codon usage occur. Such shifts would require a large number of genes to change at a large number of sites. Such shifts may possibly occur when organisms undergo long periods of reduced selection followed by an increase in selection. For example, a prolonged population bottleneck may result in reduced selection. This, in turn, may cause the levels of codon bias to become very low even in highly expressed genes. A population expansion may follow, which may increase the efficiency of purifying selection. However, the identity of the tRNA molecules that are more highly expressed and their corresponding codons may be different from before the bottleneck, leading to a shift in codon bias.

A second possibility, which would be interesting to examine experimentally, is that shifts in codon bias may occur as a result of the insertion of a new gene into a genome to which this gene is crucial to survival. For example, consider a gene conferring antibiotic resistance that is horizontally transferred into a bacterium and whose codon usage is very different from that of its new host. If the new host needs this gene to be expressed at very high levels to survive, there might be very strong selective pressure for increasing the expression of tRNA genes corresponding to those used most often by the resistance gene. Following the increase in expression of these tRNA genes, the other genes in the genome might evolve to match the new pattern of tRNA gene expression. This process might generate a genome-wide shift in codon bias.

Not yet known is how often codon bias shifts occur and how long a period of reduced or shifted selection is needed for codon bias to be erased and later changed. The ever-increasing number of sequenced genomes may help in addressing these questions.

What Is the Exact Nature of Selection on Codon Bias?

The main reason for selection on codon bias may be that the increased use of major codons leads to more efficient and more accurate translation. However, the exact relative contribution of selection for efficiency and for accuracy of translation remains unclear. Although most genes seem to be under selection to increase the use of preferred codons, some have been found to be under selection in the opposite direction (35, 42). What drives selection in these genes has not been determined. Although unlikely to explain the entire phenomenon, the use of unpreferred codons may possibly be selected for in cases in which there is a need to stall translation. For example, it may be advantageous to include rare codons in intradomain regions, because this may allow for slowdown in translation, which may then allow for better domain folding.

In addition to selection for increased efficiency and accuracy of translation, additional

selective pressures are likely to act on synonymous sites. For example, the structure of mRNA molecules is determined by their sequence, and in order to maintain their structure some properties of their sequence may be under selection. In addition, some regulatory elements, such as transcription factor binding sites, RNA localization elements, translation initiation sites, and splicing signals, may be contained within coding sequences and affected by selection.

Furthermore, specific codons may be selected for or against for reasons other than their effect on the efficiency and accuracy of translation. For example, in the immune system, B cells need to generate varied yet functional clones of their V genes, which undergo high rates of mutation. To face the opposing demands of diversification and maintenance of functional integrity, these genes use different codons in their complementarity-determining regions (CDR) than in their framework (FW) regions. Since in these proteins the CDRs need to show high variability and the FW regions are important for maintaining functional stability, the V genes evolved to overexpress codons prone to amino acid changes in their CDRs relative to their FW regions (23, 45).

Some suggestions have been made that codon bias may be a mechanism by which levels of gene expression are regulated (7). In other words, it may be that selection would be applied for a gene to have a specific level of codon bias in order for it to have a specific level of expression. This reasoning might address why in some cases selection would act in favor of unpreferred codons, but it does not provide a likely general explanation for codon bias. First, codon bias could only determine gene expression levels if elongation rather than initiation is the rate-limiting step of translation. Some evidence indicates that this is not the case (7). In addition, in a recent study Lu et al. (32) used a new method of large-scale absolute protein expression measurements to estimate the relative contribution of regulation at the levels of transcription and translation to final protein levels. They found that in yeast over 70% of gene expression regulation occurs at the level of transcription (32). Thus the contribution of codon bias to expression regulation appears to be, at best, secondary.

CONCLUDING REMARKS

While it is now clear that changes to synonymous sites are not neutral and that codon usage is affected by selection, many questions remain regarding the relationship between selection and codon bias.

Selection for the maintenance of codon bias has been shown to be weak. However, still unresolved is whether selection on silent sites is constant and whether this selection remains weak once a gene's codon bias is perturbed to a level much lower or much higher than its equilibrium. In fact, some evidence indicates that even a small number of mutations from preferred to unpreferred codons may result in significant phenotypic consequences.

FUTURE ISSUES

1. Is the strength of selection on codon bias constant?
2. What determines the choice of major codons?
3. How do shifts in codon bias between different organisms take place?
4. How much of codon bias is determined by selection on efficiency and accuracy of translation and how much by additional and sometimes even contradictory selective pressures?

Despite much interest in understanding the evolution of codon bias, these questions remain largely unanswered. We believe that the rise of genomics may help shed some light on these

questions through a combination of computational analyses of genomic datasets and high-throughput experimental studies.

DISCLOSURE STATEMENT

The authors are not aware of any biases that might be perceived as affecting the objectivity of this review.

ACKNOWLEDGMENTS

The authors thank Aia Hershberg, Anna-Sophie Fiston-Lavier, Gila Lithwick, Hanah Margalit, Josefa Gonzalez Perez, and members of the Petrov laboratory for their generous assistance. R.H. was funded by an EMBO long-term fellowship. The work was also supported by the National Institutes of Health (GM077368) grant to D.A.P.

LITERATURE CITED

1. Akashi H. 1994. Synonymous codon usage in *Drosophila melanogaster*: natural selection and translational accuracy. *Genetics* 136:927–35
2. Akashi H. 1995. Inferring weak selection from patterns of polymorphism and divergence at "silent" sites in *Drosophila* DNA. *Genetics* 139:1067–76
3. Akashi H, Kliman RM, Eyre-Walker A. 1998. Mutation pressure, natural selection, and the evolution of base composition in *Drosophila*. *Genetica* 102–103:49–60
4. Akashi H, Schaeffer SW. 1997. Natural selection and the frequency distributions of "silent" DNA polymorphism in *Drosophila*. *Genetics* 146:295–307
5. Andersson GE, Sharp PM. 1996. Codon usage in the *Mycobacterium tuberculosis* complex. *Microbiology* 142(Part 4):915–25
6. Andersson SG, Sharp PM. 1996. Codon usage and base composition in *Rickettsia prowazekii*. *J. Mol. Evol.* 42:525–36
7. Bulmer M. 1991. The selection-mutation-drift theory of synonymous codon usage. *Genetics* 129:897–907
8. Carlini DB. 2004. Experimental reduction of codon bias in the *Drosophila* alcohol dehydrogenase gene results in decreased ethanol tolerance of adult flies. *J. Evol. Biol.* 17:779–85
9. Carlini DB, Stephan W. 2003. In vivo introduction of unpreferred synonymous codons into the *Drosophila* Adh gene results in reduced levels of ADH protein. *Genetics* 163:239–43
10. Castillo-Davis CI, Hartl DL. 2002. Genome evolution and developmental constraint in *Caenorhabditis elegans*. *Mol. Biol. Evol.* 19:728–35
11. Chen SL, Lee W, Hottes AK, Shapiro L, McAdams HH. 2004. Codon usage between genomes is constrained by genome-wide mutational processes. *Proc. Natl. Acad. Sci. USA* 101:3480–85
12. Curran JF, Yarus M. 1989. Rates of aminoacyl-tRNA selection at 29 sense codons in vivo. *J. Mol. Biol.* 209:65–77
13. DuMont VB, Fay JC, Calabrese PP, Aquadro CF. 2004. DNA variability and divergence at the notch locus in *Drosophila melanogaster* and *D. simulans*: a case of accelerated synonymous site divergence. *Genetics* 167:171–85
14. Duret L. 2002. Evolution of synonymous codon usage in metazoans. *Curr. Opin. Genet. Dev.* 12:640–49
15. Duret L, Mouchiroud D. 1999. Expression pattern and, surprisingly, gene length shape codon usage in *Caenorhabditis*, *Drosophila*, and *Arabidopsis*. *Proc. Natl. Acad. Sci. USA* 96:4482–87
16. Francino MP, Ochman H. 2001. Deamination as the basis of strand-asymmetric evolution in transcribed *Escherichia coli* sequences. *Mol. Biol. Evol.* 18:1147–50
17. Fraser HB, Hirsh AE, Wall DP, Eisen MB. 2004. Coevolution of gene expression among interacting proteins. *Proc. Natl. Acad. Sci. USA* 101:9033–38

18. Ghaemmaghami S, Huh WK, Bower K, Howson RW, Belle A, et al. 2003. Global analysis of protein expression in yeast. *Nature* 425:737–41
19. Goetz RM, Fuglsang A. 2005. Correlation of codon bias measures with mRNA levels: analysis of transcriptome data from *Escherichia coli*. *Biochem. Biophys. Res. Commun.* 327:4–7
20. Gouy M, Gautier C. 1982. Codon usage in bacteria: correlation with gene expressivity. *Nucleic Acids Res.* 10:7055–74
21. Grantham R, Gautier C, Gouy M, Mercier R, Pave A. 1980. Codon catalog usage and the genome hypothesis. *Nucleic Acids Res.* 8:r49–r62
22. Hartl DL, Clarck AG. 2006. *Principles of Population Genetics*. Sunderland: Sinauer
23. Hershberg U, Shlomchik MJ. 2006. Differences in potential for amino acid change after mutation reveals distinct strategies for kappa and lambda light-chain variation. *Proc. Natl. Acad. Sci. USA* 103:15963–68
24. Ikemura T. 1985. Codon usage and tRNA content in unicellular and multicellular organisms. *Mol. Biol. Evol.* 2:13–34
25. Kanaya S, Kinouchi M, Abe T, Kudo Y, Yamada Y, et al. 2001. Analysis of codon usage diversity of bacterial genes with a self-organizing map (SOM): characterization of horizontally transferred genes with emphasis on the *E. coli* O157 genome. *Gene* 276:89–99
26. Kanaya S, Yamada Y, Kinouchi M, Kudo Y, Ikemura T. 2001. Codon usage and tRNA genes in eukaryotes: correlation of codon usage diversity with translation efficiency and with CG-dinucleotide usage as assessed by multivariate analysis. *J. Mol. Evol.* 53:290–98
27. Kanaya S, Yamada Y, Kudo Y, Ikemura T. 1999. Studies of codon usage and tRNA genes of 18 unicellular organisms and quantification of *Bacillus subtilis* tRNAs: gene expression level and species-specific diversity of codon usage based on multivariate analysis. *Gene* 238:143–55
28. Knight RD, Freeland SJ, Landweber LF. 2001. A simple model based on mutation and selection explains trends in codon and amino-acid usage and GC composition within and across genomes. *Genome Biol.* 2:RESEARCH0010
29. Lafay B, Atherton JC, Sharp PM. 2000. Absence of translationally selected synonymous codon usage bias in *Helicobacter pylori*. *Microbiology* 146(Part 4):851–60
30. Lafay B, Lloyd AT, McLean MJ, Devine KM, Sharp PM, Wolfe KH. 1999. Proteome composition and codon usage in spirochaetes: species-specific and DNA strand-specific mutational biases. *Nucleic Acids Res.* 27:1642–49
31. Lithwick G, Margalit H. 2005. Relative predicted protein levels of functionally associated proteins are conserved across organisms. *Nucleic Acids Res.* 33:1051–57
32. Lu P, Vogel C, Wang R, Yao X, Marcotte EM. 2007. Absolute protein expression profiling estimates the relative contributions of transcriptional and translational regulation. *Nat. Biotechnol.* 25:117–24
33. McDonald JH, Kreitman M. 1991. Adaptive protein evolution at the Adh locus in *Drosophila*. *Nature* 351:652–54
34. McVean GA, Vieira J. 2001. Inferring parameters of mutation, selection and demography from patterns of synonymous site evolution in *Drosophila*. *Genetics* 157:245–57
35. Nielsen R, Bauer DuMont VL, Hubisz MJ, Aquadro CF. 2007. Maximum likelihood estimation of ancestral codon usage bias parameters in *Drosophila*. *Mol. Biol. Evol.* 24:228–35
36. Precup J, Parker J. 1987. Missense misreading of asparagine codons as a function of codon identity and context. *J. Biol. Chem.* 262:11351–55
37. Sawyer SA, Hartl DL. 1992. Population genetics of polymorphism and divergence. *Genetics* 132:1161–76
38. Sharp PM, Bailes E, Grocock RJ, Peden JF, Sockett RE. 2005. Variation in the strength of selected codon usage bias among bacteria. *Nucleic Acids Res.* 33:1141–53
39. Sharp PM, Cowe E, Higgins DG, Shields DC, Wolfe KH, Wright F. 1988. Codon usage patterns in *Escherichia coli*, *Bacillus subtilis*, *Saccharomyces cerevisiae*, *Schizosaccharomyces pombe*, *Drosophila melanogaster* and *Homo sapiens*; a review of the considerable within-species diversity. *Nucleic Acids Res.* 16:8207–11
40. Shields DC, Sharp PM. 1987. Synonymous codon usage in *Bacillus subtilis* reflects both translational selection and mutational biases. *Nucleic Acids Res.* 15:8023–40
41. Shields DC, Sharp PM, Higgins DG, Wright F. 1988. "Silent" sites in *Drosophila* genes are not neutral: evidence of selection among synonymous codons. *Mol. Biol. Evol.* 5:704–16

42. Singh ND, Bauer DuMont VL, Hubisz MJ, Nielsen R, Aquadro CF. 2007. Patterns of mutation and selection at synonymous sites in *Drosophila*. *Mol. Biol. Evol.* 24:2687–97
42a. Singh ND, Davis JC, Petrov DA. 2005. X-linked genes evolve higher codon bias in *Drosophila* and *Caenorhabditis*. *Genetics* 171:145–55
43. Sorensen MA, Kurland CG, Pedersen S. 1989. Codon usage determines translation rate in *Escherichia coli*. *J. Mol. Biol.* 207:365–77
44. Stoletzki N, Eyre-Walker A. 2007. Synonymous codon usage in *Escherichia coli*: selection for translational accuracy. *Mol. Biol. Evol.* 24:374–81
45. Wagner SD, Milstein C, Neuberger MS. 1995. Codon bias targets mutation. *Nature* 376:732
46. Yamao F, Andachi Y, Muto A, Ikemura T, Osawa S. 1991. Levels of tRNAs in bacterial cells as affected by amino acid usage in proteins. *Nucleic Acids Res.* 19:6119–22
47. Yang Z. 1997. PAML: a program package for phylogenetic analysis by maximum likelihood. *Comput. Appl. Biosci.* 13:555–56
48. Yang Z. 2007. PAML 4: phylogenetic analysis by maximum likelihood. *Mol. Biol. Evol.* 24:1586–91
49. Yang Z, Nielsen R. 2008. Mutation-selection models of codon substitution and their use to estimate selective strengths on codon usage. *Mol. Biol. Evol.* 25:568–79

Annual Review of Genetics

Volume 42, 2008

Contents

Errata

An online log of corrections to *Annual Review of Genetics* articles may be found at http://genet.annualreviews.org/errata.shtml